基于深度学习的中文微博作者身份识别研究 *

徐晓霖，蔡满春，芦天亮

(中国人民公安大学 信息技术与网络安全学院，北京 102623)

摘 要：作者身份识别一直在公安行业和文检工作中起着重要的作用。现有的作者语言风格建模过程繁琐、文本特征工程没有普适性。针对此问题，在无须专家进行特征建模的情况下，提出 CABLSTM 中文微博作者身份识别模型，并在公开微博语料集测试该模型准确度。该模型为最大化的提取短文本特征，融合 Attention 机制于 CNN 中并去除池化层，通过双向 LSTM 以获取上下文相关信息，身份识别结果通过 Softmax 层进行输出。实验结果表明，该模型在进行中文微博作者身份识别任务中与传统机器学习算法以及 TextCNN 和 LSTM 算法相对比，在准确率、召回率、F 值方面都有一定的提升。

关键词：作者身份识别；LSTM；CNN；特征自动提取

中图分类号：TP391.72 doi: 10.19734/j.issn.1001-3695.2018.05.0486

Basics depth academic learning Chinese fumiohiro writer authorship identification research

Xu Xiaolin, Cai Manchun, Lu Tianliang

(*School of Information Technology& Network Security*, *People's Public Security University of China*, *Beijing* 102623, *China*)

Abstract: Author identification has always plays an important role in the public security and literary inspection work. Texts feature extraction is cumbersome and not universal. To solve this problem, the CABLSTM Chinese microblog author identification model is proposed without expert feature modeling, and the accuracy of the model is tested in the open microblog corpus. This model maximizes the extraction of short text features, fuses the Attention mechanism in the CNN and removes the pooling layer, and obtains context-related information through the bidirectional LSTM. The identity recognition result is output through the Softmax layer. Experimental results show that the model has a certain improvement in accuracy, recall rate, and F value in comparison with traditional machine learning algorithms and TextCNN and LSTM algorithms in the identification task of Chinese microblog authors.

Key words: author identification; LSTM; CNN; automatic feature extraction

0 引言

文本作者身份识别是文检言语分析中的类别，研究属于应用语言学和计算机科学的交叉领域，其主要思路是将文本中隐含的作者无意识写作习惯通过某些可以量化的特征表现出来，凸显作品的文体学特征和写作风格，以此确定匿名文本的作者。

文检工作中的言语分析就是根据文本的写作风格从而确定匿名文本作者。公安工作中有害信息作者鉴定也可以基于文本对嫌疑人员进行判断。文本作者身份识别为上述两种提供一定的分析支持。

前人研究的文本作者身份识别，大多集中在长文本。人们从一元论文本特征到多元论文本特征再到多层次文本特征，不断提高了对文本特性的抽取，更深力度更加抽象地进行文本特征抽取建模，以提高文本作者身份是别的准确性。但随着网络的急速发展，网络文本大量涌现，邮件、博客、微博、评论等等短文本大量存在但长文本作者身份识别方法并不能完全适用于短文本。现阶段对于短文本的研究较少，只有祁瑞华等人[1,2]针对微博短文本通过词汇、句子、依存关系、特殊符号等多方面特征提取进行特征建模，实现了基于短文本的文本作者身份识别。但是这种方法并不能对所有的短文本进行统一的特征提取，不同的短文本需要不同的特征提取方式，且微博中的特殊符号等大大增加了判断的准确率，这在普通短文本中并不具有。

大量微博内容是少于 140 个字的，很难在如此短的文本中提取文本特征，但微博的发言往往是作者很随性的，更能代表出作者的语言风格。现阶段根据多种文本特征和微博特有特征的文本特征提取方法虽然取得了很好的效果，但都是对于某一

收稿日期：2018-05-29；**修回日期：**2018-07-11 **基金项目：**国家重点研发计划重点专项资助项目（2017YFB0802804）；国家自然科学基金资助项目（61602489）；中国人民公安大学 2018 年基本科研业务费科研机构项目（2018JKF504）

作者简介：徐晓霖（1994-），男，山东济南人，硕士，主要研究方向为网络安全与执法（1326227533@qq.com）；蔡满春（1972-），男，副教授，博士，主要研究方向为网络安全、密码学；芦天亮（1985-），男，副教授，博士，主要研究方向为网络安全、恶意代码分析与检测.

特定的短文本，并且无法避免由专家学家进行人工特征建模的过程，为此本文尝试提出基于深度学习的中文微博作者身份识别模型，通过深度学习对短文本进行文本自动特征提取，去掉专家特征建模过程，并在公开微博与语料上测试共有效性。

1　中文微博作者身份识别模型

深度学习具有能够自主学习并提取特征的特性，因此本文想通过深度学习实现文本作者身份识别。卷积神经网络（CNN）和长短时记忆网络（LSTM）是深度学习中较为流行的分类模型。卷积神经网络（CNN）有着类似于 n-gram 的效果，能够提取文本特征，并且通过多卷积层进行更加深入的挖掘，因此考虑将 CNN 与 Attention 进行结合扩大及加强其特征提取的特效作为短文本特征提取器。长短时记忆网络（LSTM）是对时序数据进行特征提取，能够有效获取上下文信息，因此考虑将其与分类器结合作为输出器。

CABLSTM 中文微博作者身份识别模型的建立流程如图 1 所示。

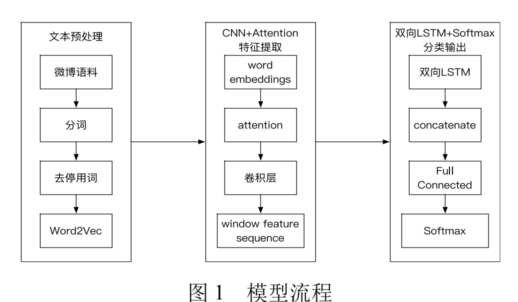


图 1　模型流程

Fig.1　Mode flowchat

1）文本预处理

将 40 G 微博语料由添加微博热词为用户自定义词典的 NLPIR 进行分词，去停用词后输入到 Word2Vec[3, 4]中产生词向量。

2）CNN+Attention 特征提取

将句子分词后组成的词向量矩阵进行 attention 后获得双通道 conv input（sentence word enbeddings，attention feature map）作为 CNN[5, 6]卷积层的输入，卷积后按位组合得到 window feature sequence。即文本特征向量。

3）双向 LSTM+Softmax 分类输出

将 window feature sequence 作为双向 LSTM[7, 8]的输入，得到两个一维向量，通过 concatenate 层进行拼接，最后通过全连接层和 Softmax 进行分类输出

1.1　文本预处理

1）微博爬取

本文所需要的数据分为两类，一类为建立词向量所需要的大量微博数据；一类为实验所需的以作者为标签的分类微博。由于建立词向量需要大量的数据，所以本文采用的是在 CSDN 上获得的 40 G 微博数据。实验数据通过 python 的 request 包和正则表达式进行爬取。首先人工选择符合要求且筛选出发博量超过 1000 条的候选人，进行 10 人次的数据爬取，共 10000 条。

2）文本分词

首先对训练语料进行中文分词。现阶段开放 python 接口且较流行的分词工具为 Jieba、NLPIR、LTP。本文选择进行微博语料分词实验后准确率最高的 NLPIR 分词工具。

3）词向量生成

文本分词结束后，去掉停用词后采用 Word2vec[4]的 CBOW 模型进行词向量的建立。输入层为单词 x 周围的 n-1 个词向量，将 n-1 个词向量相加输入到隐藏层；然后从根节点开始，映射层的值需要沿着 Huffman 树不断的进行 logistic 分类，并且不断修正各中间向量和词向量；最后输出单词 x 的词向量。

1.2　CNN+Attention 机制进行文本特征提取

由于微博数据过于短小，所以需要尽可能地对文本进行更抽象和高级的特征特征表示，才能够更行文本特征建模。CNN 可以进行卷积操作，考虑到卷积的效果且 CNN 具有 n-gram 特征提取的能力，所以用 CNN 能够更好地对微博短文本进行文本特征提取。为了能够更深度挖掘特征，对 CNN 进行改进。

a）去掉 CNN 中的 Max-Pooling 层。虽然池化层可以降低输出的向量维度，但同时也丢失了部分特征，因此在进行特征提取时将 Max-Pooling 层去掉，以充分发挥 CNN 卷积提取特征的效果。

b）在 CNN 进行卷积前加入 Attention 层。传统的 CNN 通过每个单通道处理一个句子，然后学习句子表达，最后一起输入到分类器中。该模型在输入分类器前句对间没有相互联系，只能学习到局部特征，通过 Attention[9, 10]机制，将句子 s1 与句子 s2 构建 attention 矩阵，s1 通过与 attention 矩阵相乘得到 attention feature map，卷积层的输入由单通道变为双通道，将不同 cnn 通道的句对联系起来，可以学习全文特征，提高特征提取的效果。

如图 2 所示，首先计算 attention 矩阵 A，其每个元素 $A_{i,\ j}$ 代表句子 1 中第 i 个单词对句子二中第 j 个单词的 match_score，经验表明当 match_score 为 Euclidean 距离时效果很好为此，本文选用 Euclidean 距离作为 match_score 计算公式为

$$A_{i,\ j}=\frac{1}{1+|x-y|}\left(F_{0,r}[:,i],F_{1,r}[:,j]\right) \tag{1}$$

W_0 和 W_1 均为学习优化的的参数矩阵,本文使用相同的 W，即共享两个矩阵。这样 W_0 和 A 的转置想乘，W_1 和 A 想乘得到两个句子的与原句子词向量矩阵大小相同的 attention feature map。计算公式为

$$F_{0,a}=W_0\times A^T \tag{2}$$

$$F_{1,a}=W_1\times A \tag{3}$$

一个句子由其本身分词后的词向量矩阵与其 attention feature map 这两个通道作为 CNN 卷积层的输入。选择固定窗口大小的 filter 进行卷积获得 feature maps。由于需要输入到 LSTM 中，所以将每个 feature maps 的对应位置进行拼接构成 window feature sequence，并且不连接 Pooling 层以最大化的挖掘文本特征。

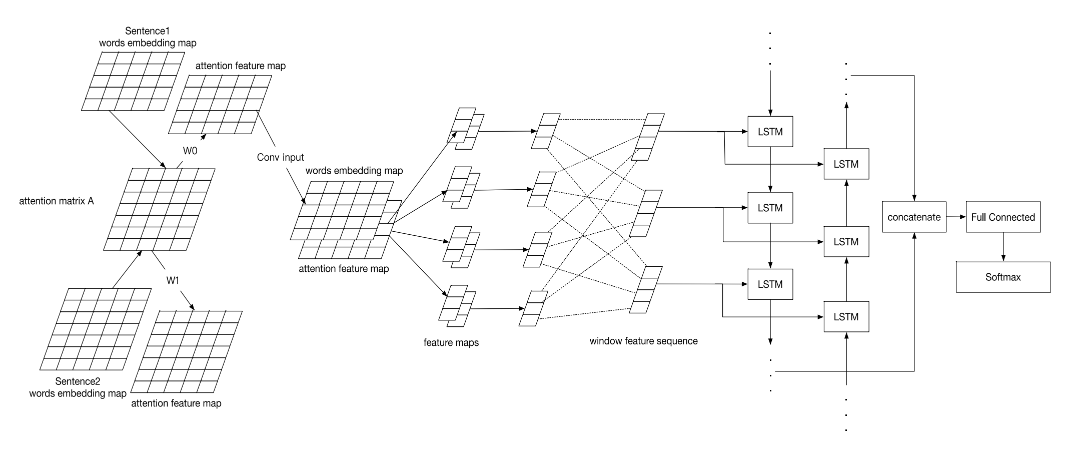


图 2 CABLSTM 模型

Fig.2 CABLSTM mode

1.3 双向 LSTM+Softmax 进行分类输出

单向 LSTM，在 t 时刻的输入 I_t 代表的是 t 时刻之前的输入信息，该信息包含了上文中的信息，但是并不包含下文中的信息；而使双向 LSTM 算法加入了反方向的 LSTM，这对于一个单元 t 时刻的输入 I_t 和 $\tilde{I}_t$ 分别代表了上文信息和下文信息。所以，为了更好地提取微博短文本的文本特征，选择使用双 LSTM+Softmax[11]进行分类输出。

如图 2 所示，CNN+Attention 进行文本特征提取将 window feature sequence 输入到双向的 LSTM 中得到两个一维向量，为了使特征更好地保留，不选择 aver 而选择 concatenate 进行两个向量的拼接，以免特征的剔除。最后，添加全连接层和 Softmax 层进行分类。

2 实验分析

2.1 实验数据来源

本实验所选择的实验数据为自己爬取的新浪微博为实验语料，收集了新浪微博 10 位公众人物的共 10 000 篇微博，每位 1 000 篇。其中语料最长的为 140 字，最短的为 45 字。采用十字交叉进行实验，在各组对照实验中，统计作者身份识别的准确率（precision）、召回率（recall）和 F-measure 的平均值评估作者身份识别性能。

2.2 实验环境

所有实验基于 Python 3.6 来实现，使用 Alineware 机器，CPU 为 i7、内存 16 GB、系统为 Linux、显卡为 gtx1070。

2.3 微博分词准确率对比实验

采用 Jieba、NLPIR 和 LTP 三种较为流行的分词工具进行准确性实验，大多以"人民日报分词语料集"作为实验数据。人民日报文本十分严格规范，口语化程度较低，网络流行语较少，文本长度较长，与微博语料有较大的差别。为了更好地选择对微博类短文本分词准确率较高的工具，本节以微博语料为数据源，对三种分词工具进行对照实验。

因为没有标准的分词后的微博语料库，所以人工对 3 000 篇爬虫的微博语料进行人工分词，以'|'为分割符。分词例图如表 1 所示。

表 1 分词例图

Table 1 Word segmentation example

一条微博数据	
原句	终极预告终于和大家见面啦~~史无前例的悉尼歌剧院打斗。
人工分词	终极|预告|终于|和|大家|见面|啦|史无前例|的|悉尼|歌剧院|打斗

实验流程如图 3 所示。

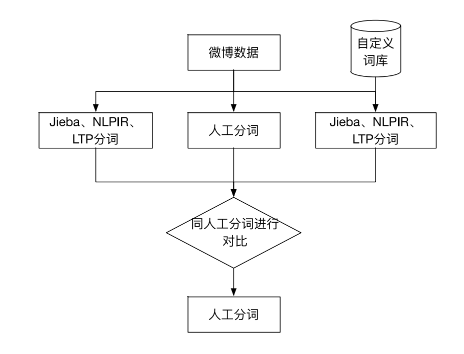


图 3 分词实验流程

Fig.3 Word segmentation experiment flowchat

实验分三组数据进行比较。数据 1：人工分词；数据 2：Jieba、NLPIR、LTP 分词；数据 3：加入用户自定义词典后的 Jieba、NLPIR、LTP 分词。用户自定义词典由近五年微博热词、微博网络语组成。准确率由 3 000 条微博人工分词对比结果。时间由 100 000 条微博测试得出。

表 2 分词工具实验结果

Table 2 Word segmentation experimental result

是否添加外界词库	工具	准确率	时间
未添加用户自定词词典	Jieba	91%	149.2seconds
	NLPIR	96%	53.5 seconds
	LTP	94%	175.7seconds
添加用户自定义词典	Jieba	94%	162.4seconds
	NLPIR	98%	63.2seconds
	LTP	97%	187.4seconds

实验结果如表 2 所示。可以看出：a）总体上，三种算法在中文微博分词的准确率都达到了 90%以上，证明了三种分词工

具都是十分有效的；b）从是否添加用户自定义词典来看，添加用户自定义词典后三种分词算法的准确率都有了一定的提高；c）从算法性能来看，NLPIR 中国科学院的分词工具无论是否添加用户自定义词典都是分词准确率最高的，加入用户自定义词典后准确率可高达 98%；d）从时间来看，三种分词工具中 NLPIR 中国科学院分词工具的时间最短，时间约只有其他两种算法三分之一。因此，从准确率和时间消耗来分析，NLPIR 中国科学院分词工具在加入用户自定义词典后有着最高的分词准确率和最低的时间消耗，在实验进行词向量建立时能够更好地完成任务，且具有更好的分词准确率，可以使模型的效果得到一定的提升。

2.4　中文微博作者身份识别算法对比实验

为了验证 CABLSTM 模型在中文微博作者身份识别算法，本文采用作者身份识别中常用的准确率（P）、召回率（R）和 F-measure 值作为指标来测量其有效性及优越性。通过 F1 值，能够结合准确率及召回率更加客观地反映出该模型的综合水平。

首先验证 CABLSTM 模型在中文微博作者身份识别中的有效性，采用 SVM、决策树 C4.5、TextCNN、LSTM 和 CABLSTM 五种算法。在 SVM 和决策树 C4.5 算法中，中文微博特征集由词汇频率特征、标点数量特征、功能词次数特征、词性标注特征组成。实验结果如表 3 所示。可以看出：a）总体上，五种算法在中文微博作者身份识别任务上准确率、召回率和 F-measure 均达到了 70%以上，每位作者的准确率、召回率和 F-measure 都达到了 69%以上；b）从算法性能来看，TextCNN 和 LSTM 在准确率、召回率和 F-measure 方面和人工特征建模的传统机器学习 SVM 和 C4.5 相差不大。改进后的 CABLSTM 模型在准确率、召回率和 F-measure 方面相对另外四种算法都有一定程度的提高。具体来说本文模型可以更加深力度的挖掘短文的文本特征，为分类提供更好的特征模型。利用 CNN 加 Attention 机制能够提高深度学习对文本的学习提取力度；利用双向 LSTM 加 Softmax 能够更好地学习特征并进行分类。

CABLSTM 算法在完成中文微博作者身份识别任务中，与传统的机器学习算法 SVM 和 C4.5 算法相比较而言，去掉了以词汇频率特征、标点数量特征、功能词次数特征、词性标注特征为特征集的文本特征建模过程，在提高准确率的同时减少了人工的参与，提高了效率，降低了人工特征建模的难度；与 TextCNN 和 LSTM 相比，在准确率、召回率和 F 值得到了一定的提高。所以 CABLSTM 模型可以更好地应用于中文微博作者身份识别，为公安行业有害信息作者识别和文检工作提供一定的理论支持和技术支持。

表 3　实验结果

Table 3　Experimental result

算法		作者 1	作者 2	作者 3	作者 4	作者 5	作者 6	作者 7	作者 8	作者 9	作者 10	加权平均
SVM	P	0.71	0.75	0.72	0.74	0.8	0.69	0.72	0.73	0.75	0.79	0.74
	R	0.77	0.7	0.71	0.7	0.75	0.82	0.63	0.87	0.82	0.7	0.747
	F	0.74	0.79	0.71	0.71	0.77	0.75	0.67	0.79	0.78	0.74	0.745
C4.5	P	0.8	0.79	0.82	0.84	0.79	0.83	0.85	0.77	0.75	0.8	0.804
	R	0.8	0.81	0.8	0.82	0.81	0.8	0.75	0.77	0.79	0.78	0.793
	F	0.8	0.79	0.81	0.83	0.79	0.81	0.79	0.77	0.76	0.79	0.794
TextCNN	P	1	0.95	0.82	0.74	0.89	0.7	0.95	0.99	0.82	0.57	0.843
	R	0.99	0.91	0.92	0.41	0.79	0.81	0.9	0.98	0.92	0.75	0.838
	F	0.99	0.93	0.87	0.52	0.83	0.76	0.93	0.99	0.87	0.65	0.834
LSTM	P	0.99	0.73	0.66	0.49	0.87	0.69	0.94	0.98	0.46	0.47	0.728
	R	0.97	0.87	0.63	0.44	0.71	0.82	0.79	0.93	0.96	0.72	0.784
	F	0.98	0.79	0.65	0.47	0.78	0.75	0.86	0.96	0.62	0.57	0.743
CABLSTM	P	1.0	0.92	0.97	0.87	0.94	0.94	1.0	1.0	1.0	0.82	0.964
	R	0.99	0.98	0．97	0.71	0.94	0.92	0.98	0.99	0.98	0.89	0.935
	F	0.99	0.95	0.97	0.78	0.94	0.93	0.99	0.99	0.99	0.88	0.941

3　结束语

本文拓展了作者身份识别研究的理论框架和应用范围，考虑传统长文本和网络短文本在文本特征提取上的差异，研究前人对于短文本特征建模的改进。针对现阶段中文微博作者身份识别必须人工进行文本特征建模的现状，本文提出了基于深度学习算法的 CABLSTM 模型进行微博文本特征提取并进行文本作者身份识别，去掉了文本作者身份识别中必须人工特征建模的过程，减少了人工投入，提高了效率。这样在公安行业用拥有重点人群和其发表言论库时，可以使用该模型对无法判别作者的有害言论进行一定的分析，从而为公安行业和文检行业的作者识别提供一定的理论和技术支持。下一步的研究计划是研究如何在作者数量较多的中文微博语料上提高中文微博作者身份识别的准确率。



参考文献：

[1] 祁瑞华，杨德礼，郭旭，等. 基于多层面文体特征的博客作者身份识别研究 [J]. 情报学报，2015, 34 (6): 628-634. (Qi Ruihua, Yang Deli, Guo Xu. Blogger identification based on multidimensional stylistic features [J]. Journal of the China Society for Scientific and Technical Information, 2015, 34 (6): 628-634.)

[2] 祁瑞华，郭旭，刘彩虹. 中文微博作者身份识别研究 [J]. 情报学报，2017, 36 (1): 72-78. (Qi Ruihua, Guo Xu, Liu Caihong. Authorship attribution of Chinese microblog [J]. Journal of the China Society for Scientific and Technical Information, 2017, 36 (1): 72-78.)

[3] Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26: 3111-3119.

[4] Zhang Dongwen, Xu Hua, Su Zengcai, *et al.* Chinese comments sentiment classification based on word2vec and SVM perf [J]. Expert Systems with Applications, 2015, 42 (4): 1857-1863.

[5] Roska T, Chua L O. The CNN universal machine: an analogic array computer [J]. IEEE Trans on Circuits & Systems II Analog & Digital Signal Processing, 2015, 40 (3): 163-173.

[6] Chua L O, Roska T. The CNN paradigm [J]. IEEE Trans on Circuits & Systems I Fundamental Theory & Applications, 1993, 40 (3): 147-156.

[7] Gers F A, Schmidhuber J, Cummins F. Learning to forget: continual prediction with LSTM [J]. Neural Computation, 2000, 12 (10): 2451-2471.

[8] Graves A. Supervised sequence labelling with recurrent neural networks [J]. Studies in Computational Intelligence, 2008, 385.

[9] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. Computer Science, 2014.

[10] Xu K, Ba J, Kiros R, *et al.* Show, attend and tell: neural image caption generation with visual attention [J]. Computer Science, 2015: 2048-2057.

[11] Salakhutdinov R, Hinton G E. Replicated softmax: an undirected topic model [C]// Proc of International Conference on Neural Information Processing Systems. [S. l.] : Curran Associates Inc, 2010: 1607-1614.

[12] Roska T, Chua L O. The CNN universal machine: an analogic array computer [J]. IEEE Trans on Circuits & Systems II Analog & Digital Signal Processing, 2015, 40 (3): 163-173.